Robust Hierarchical k-Center Clustering

Silvio Lattanzi
Google Research, New York
silviol@google.com

Stefano Leonardi*
Sapienza University of Rome
leonardi@dis.uniroma1.it

Vahab Mirrokni
Google Research, New York
mirrokni@google.com

Ilya Razenshteyn†
MIT
ilyaraz@mit.edu

ABSTRACT

One of the most popular and widely used methods for data clustering is hierarchical clustering. This clustering technique has proved useful to reveal interesting structure in the data in several applications ranging from computational biology to computer vision. Robustness is an important feature of a clustering technique if we require the clustering to be stable against small perturbations in the input data. In most applications, getting a clustering output that is robust against adversarial outliers or stochastic noise is a necessary condition for the applicability and effectiveness of the clustering technique. This is even more critical in hierarchical clustering where a small change at the bottom of the hierarchy may propagate all the way through to the top.

Despite all the previous work [2, 3, 6, 8], our theoretical understanding of robust hierarchical clustering is still limited and several hierarchical clustering algorithms are not known to satisfy such robustness properties. In this paper, we study the limits of robust hierarchical k-center clustering by introducing the concept of *universal hierarchical clustering* and provide (almost) tight lower and upper bounds for the robust hierarchical k-center clustering problem with outliers and variants of the stochastic clustering problem. Most importantly we present a constant-factor approximation for optimal hierarchical k-center with at most z outliers using a universal set of at most $O(z^2)$ set of outliers and show that this result is tight. Moreover we show the necessity of using a universal set of outliers in order to compute an approximately optimal hierarchical k-center with a different set of outliers for each k.

1. INTRODUCTION

As the amount of data available in many different domains (social sciences, text classification, bioinformatics, image processing) is increasing every day, the development of computationally efficient data clustering techniques that are also effective, accurate, and robust to noise has become increasingly important. As one of the most popular and widely used methods for data clustering, hierarchical clustering aims to describe the structure of the data at different scale levels. This clustering technique is applied as a standard tool by statisticians, computer scientists, bio-scientists and more recently data scientists.

Hierarchical clustering provides partitions of the input data at different levels where every partition is obtained by refining a partition at a higher level. The structure of the solution of a hierarchical clustering algorithm is conveniently represented by a tree. More specifically, the top partitioning is formed by the whole set of input points, the partitioning of level k is obtained by refining one of the clusters of the partition at level $k-1$, and the partitioning at the lowest level is the one with each of the points in a separate cluster. The quality of a partitioning is mathematically described by some measures of the internal cohesiveness of the clustering, every measure giving rise to a different clustering problem. In this paper we consider the k-center clustering problem, i.e., the problem of partitioning data into k clusters while minimizing the maximum radius of a cluster.

The k-center clustering problem is NP-hard, and admits simple aproximation algorithms with a ratio of 2 [11, 14]. An approximation algorithm for hierarchical k-center clustering problem outputs a tree of partitions that contains, for every integer k, a k-cluster partitioning that is a good approximation of the optimum k-center clustering. For hierarchical k-center clustering, Dasgupta and Long [8] presented an algorithm that for every k induces a k-clustering with maximum radius at most 8 times the optimal k-center clustering. A wealth of methods for computing a hierarchical clustering has been proposed in the literature (see for instance [3, 6, 8, 10, 16, 17].) These methods are either based on a top-down recursive partitioning or more often based on a bottom-up agglomerative method that merges the two closest clusters according to some measure.

Most of the standard hierarchical clustering algorithms are not tolerant to noise [20] meaning that the output solution is very sensible to size and position of the outliers. In other words, the insertion of a few outliers in the metric space may determine a complete alteration of the clustering structure, thus leaving the question of whether the computed clustering is meaningful at all. In order to deal with issue, several practical and theoretical algorithms have been proposed. Some well-known practical techniques are the Ward's method [21], the Wishart's method [22], and CURE [13]. These algorithms however do not have a theoretical guarantee. Various attempts for a theoretical analysis of these problems have also been made, e.g., Balcan et. al [2] studied a certain type of separation property, and presented an agglomerative procedure that clusters instances that possess this property.

*Work partially done while at Google Research NY. Partially supported from Google Focused Research Award "Algorithms for Large-scale Data Analysis", EU projects FET MULTIPLEX 317532 and ERC PAAI 259515.
†Work partially done while at Google Research NY.

ITCS'15, January 11–13, 2015, Rehovot, Israel.

http://dx.doi.org/10.1145/2688073.2688104.

In this paper, we study a variant of the robust hierarchical k-center clustering. Specifically, we seek hierarchical clustering algorithms that are tolerant against *outliers*, i.e., we aim to find, for each k, a solution to the k-center problem that is approximately optimal for a large portion of the data points excluding a small number of outliers (considered as noise in the data). In particular, our goal is to compute a k-center hierarchical clustering that is robust against outliers using a small set of *universal outliers*, i.e., a set of outliers whose removal allows to find an approximately optimal solution for the k-center problem for all values of k. Although dealing with outliers has been extensively considered already in the context of k-center and other clustering problems[4, 5, 19], the question of finding a universal set of outliers for all k has not been formally studied. In fact, in our study, we will show that a universal set of outliers is needed if we want to obtain a good approximation for all values of k, since an outlier used for one partition of the hierarchy may be used as a center for another partition. In this paper, we aim to study this problem by answering the following question: Does there exist a set of *universal outliers* whose removal allows to compute a k-center hierarchical clustering on the remaining points that is approximately optimal for all values of k?

We also investigate another aspect of robustness related to clustering. In applications with large-scale data it is often impossible to observe the whole input data, but it is available some knowledge on the distribution that generates the input. It is relevant in these cases to compute a clustering that is good on average for most of the input instances generated by the distribution. Achieving a similar result requires to discard some outliers, for instance those points that have a big influence on the solution but appear in the input data with small probability. In this work, we address for the first time the problem of computing an *hierarchical stochastic k-center clustering with outliers* that provides, on average, a good solution for (almost) any sample of the metric space.

Our Contributions. We study the robust hierarchical k-center clustering problem. One of our main contributions is to introduce the concept of *universal hierarchical clustering* and provide lower and upper bounds for the hierarchical clustering problem with outliers and variants of the stochastic clustering problem.

To achieve this result, we first study the existence of a set of *universal outliers*, O, that can be used for k-center solution for all k, $k = 1,\ldots,n$, to obtain a solution that compares with the optimal solution of the k-center clustering with z outliers. In particular, we present the following structural results on the set of *universal outliers*:

1. There does not exist a constant-factor approximation algorithm for all possible k using a set of universal outliers with less than $\Omega\left(z^2\right)$ points.

2. There does not exist a constant-factor approximation algorithm for the problem without deciding in advance a universal set of outliers.

We also present the following nearly matching upper bound:

3. There exists a set of universal outliers, O, with $O\left(z^2\right)$ points, such that the k-center solution computed with O as a set of outliers is an $O(1)$ approximation of the optimal k-center with z outliers for all possible k(where the z outliers can change with k).

4. The set of universal outliers can be computed in polynomial time.

The above result is somehow surprising as it implies that the total number of points in the set of *universal outliers* does not depend on n, or on the number of levels of the hierarchical clustering, but only on z.

We next consider the problem of designing a stochastic hierarchical k-center clustering with outliers that provides a good hierarchical solution for every subset of h points independently sampled from a metric space of n points.

5. There exists an $O(1)$-approximation algorithm for *stochastic hierarchical* k-center clustering on inputs of size h and $k \leq K$ that uses at most $z = \frac{2K(K-1)\ln n}{h} n$ outliers.

The above result is obtained by computing an *a-priori* solution that is used for all subsets sampled from the metric space. In order for this result to be non-trivial, K should be bounded by $O\left(\sqrt{\frac{n\times h}{\ln n}}\right)$. In our stochastic model, all points are sampled with equal probability. However, we observe that a non-uniform distribution can be simulated by repeating more points at nearby locations of the metric space. With suitable discretization, this leads only to a small loss in the approximation for the k-center problem.

2. RELATED WORK

Hierarchical or agglomerative clustering is a popular clustering techniques widely studied in data mining and machine learning [8, 15, 18].

We study the k-center clustering problem with outliers. The first work in this direction was by Charikar et al. [4] that develops a 3-approximation algorithm for the problem. If one is interested in hierarchical k-center clustering (without outliers), [8] then gives a procedure for finding a hierarchical clustering that is 8-approximated for every $k = 1,\ldots,n$. This paper is to the best of our knowledge the first theoretical work which combines both hierarchical k-center clustering and the notion of outliers. Our construction for universal outliers is somewhat similar conceptually to [19], where a streaming algorithm for k-center clustering with outliers is presented. Finally, universal stochastic optimization has been considered for network design and set cover problems in [12, 9]. The problem of universal stochastic optimization with outliers for network design has also been addressed in [1].

3. PRELIMINARIES

In this section, we give a formal definition of the hierarchical k-center with z outliers, and introduce some notation that will be used throughout the paper.

Let $\mathcal{M} = (X,d)$ be a metric space, S be a subset of X and $n = |X \setminus S|$. Let $c_1, c_2, \ldots, c_n$ be an ordering of the points in $X \setminus S$ and π be a function from $\{c_2,\ldots,c_n\} \to \{c_1,\ldots,c_n\}$ such that for every c_i we have $\pi(c_i) = c_j$, with $j < i$. Note that using the function π, we can define a tree on $c_1, c_2, \ldots, c_n$ where each node c_i points to $\pi(c_i)$ and c_1 is the root of the tree.

For every point $c_i \in X \setminus S$ with $i > 1$, we say that the node c_j is the ancestor of c_i in $c_1, c_2, \ldots, c_k$, with $1 \leq k \leq i-1$, and denote it by $a_k(c_i) = c_j$, if $c_j \in \{c_1, c_2, \ldots, c_k\}$ is the closest node in π to c_i.

Note that we can define a recursive partition of the points in $X \setminus S$, using $c_1, c_2, \ldots, c_n$ and π. More precisely, for every $1 \leq k \leq n$ we can use the first $c_1,\ldots,c_k$ points as centers of the clusters in the partition and we can assign each node c_i with $i > k$ to the cluster whose center is its ancestor. Furthermore, we can define the cost of such a partition at level k as $C(c_1, c_2, \ldots, c_k, \pi, X \setminus S) = \max_{i>k} d(c_i, a_k(c_i))$.

For $1 \leq k \leq n$, let O_k denote the optimal set of z outliers for k-center clusterings, and OPT_k^z denote the cost of the optimal solution for k-center with z outliers and with $\text{OPT}_k(X \setminus S)$ be the cost of the optimal k-center solution for the points in $X \setminus S$. In the hierarchical k-center problem with z outliers, we want to find an ordering of points in $X \setminus S$, $c_1, c_2, \ldots, c_n$, and a function π such that $C(c_1, c_2, \ldots, c_k, \pi, X \setminus S) \leq \alpha \text{OPT}_k^z$ for any $1 \leq k \leq n$ where α is minimized.

In the rest of the paper, we denote by $B(u, C)$ the ball centered in u and with radius C. Also we often refer to S as universal set of outliers.

4. OUTLIERS AND HIERARCHICAL K-CENTER

In this section, we analyze the structural properties of hierarchical k-center clustering with outliers. We know from the work of Das and Mathieu [7](Theorem 4) that no randomized algorithm can achieve an approximation better than 3/2 for the hierarchical k-center clustering problem even when we allow unbounded computational power so in the paper we focus on finding a constant factor approximation for our problem.

In particular we are able to show that it is possible to compute a hierarchical k-center clustering that uses a set of universal outliers S of size $O\left(z^2\right)$ and that gives a constant approximation for the k-center problem with z outliers for all values of k. Furthermore, we show that this result is asymptotically tight, i.e., in order to get a constant factor approximation for all k, we need to use a universal set of outliers and such a universal set of outlier has to be of size at least $\Omega\left(z^2\right)$.

In the next subsections, we first present the two structural lower bounds, and then we describe our positive result.

4.1 Lower bounds

4.1.1 Size of the set of universal outliers

Given that we cannot hope to have an approximation better than 3/2, we focus on obtaining a constant-factor approximation. Here we show that in order to get a constant-factor approximation algorithm with a set of universal outliers S, we need S to be of size $\Omega\left(z^2\right)$.

LEMMA 1. *There exists a layout of the points in a 1-dimensional euclidian space such that it is impossible to obtain a constant approximation for all possible k using a set of universal outliers with less than* $\Omega\left(z^2\right)$ *points.*

PROOF. We partition the n points in t sets $U_0, U_1, \ldots, U_{t-1}$. All the sets U_j, for j in $1 \leq j \leq t-1$, contain $j+1$ points, the set U_0 contains the remaining points. All the points in U_0 lie in $[0,1]$, in all the other sets U_j, for j in $1 \leq j \leq t-1$, there is one point in position $-\log^j(n)$ and j points in $[\log^j(n) - 1, \log^j(n)]$.

Consider the 1 center solution for the problem with z outliers. The optimal solution in this case is to select one center in U_0 and to select as outliers all points in U_{z-1}, this solution has cost $O(log^{z-2}n)$. Furthermore note that all the solutions with cost $O(log^{z-2}n)$ select U_{z-1} as set of outlier so the universal set of outliers contains U_{z-1}. The optimal solution with 2 centers and z outliers has one center in U_0 one center in the positive points of U_{z-1}, and the outliers' set is composed by the negative point in U_{z-1} and the points in U_{z-2}, this solution has cost $O(log^{z-3}n)$. Also in this case it is possible to check that there is no solution of cost $O(log^{z-3}n)$ with a different set of outliers so the universal set of outliers contains $U_{z-1} \cup U_{z-2}$. More generally, the solution with j centers and z outliers, for $1 \leq j \leq z-1$, has one center in U_0, one center between the positive points in U_{z-1}, one center between the positive points in $U_{z-2}, \ldots$, one center between the positive points in U_{z-j+1} and the set of outliers for such solution is composed by the negative points of $U_{z-1}, U_{z-2}, \ldots, U_{z-j+1}$ and all the points in U_{z-j}(refer to figure 1 for an example for $k = 1, 2, 3$ and $z = 4$). This solution has cost $O(log^{z-j-1}n)$ and one can check that it is possible to get a $O(log^{z-j-1}n)$ solution only using the described set of outliers, so the universal set of outliers contains $U_{z-1} \cup U_{z-2} \cup \cdots \cup U_{z-j}$.

Thus we have that the cardinality of the universal set of outliers U is lower bounded by

$$|U| \geq \left|\cup_{j=1}^{z-1} U_j\right| = \sum_{j=1}^{z-1} \left|U_j\right| = \sum_{j=1}^{z-1} j + 1 \in O(z^2).$$

The claim follows. □

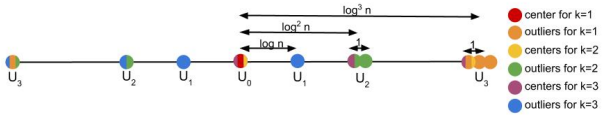

Figure 1: Visualization of the layout for $z = 4$. The different set of outliers and centers are represented using different colors. Nodes with multiple colors are outliers or centers for multiple values of k.

4.1.2 A universal set of outliers is necessary

We just proved that the size of the set of universal outliers needed to get a constant-factor approximation is at least of size $\Omega\left(z^2\right)$. In this subsection, we show that a universal set of outliers is necessary if we want to construct a robust hierarchical k-center solution. More precisely, we show that it is impossible to obtain a constant-factor approximation for the problem without deciding in advance a universal set of outliers to be excluded since the beginning of the execution of a hierarchical clustering algorithm. To prove it we exhibit an example where one of the centers in a k-center solution computed for a value of k must later be turned into an outlier if we like to maintain a constant approximation for another value of k (note that this is impossible in a hierarchical solution). Our result holds also if we are allowed to use a different set of outliers for different values of k.

LEMMA 2. *There exists a layout of the points in a 1-dimensional Euclidian space such that for $z = 4$ it is impossible to obtain an approximate solution for all possible k without using a set of universal outliers.*

PROOF. Consider the following layout of the points, point v_1 is in position $\log^2 n$, point v_2 in position $\log^2 n + 1$, point v_3 in position $-\log^2 n$, point v_4 in position $-\log^2 n - 1$, point v_5, v_6, v_7 and v_8 are respectively in positions $\log^3 n, \log^3 n + \log n, \log^3 n + 2\log n$ and $\log^3 n + 3\log n$, all the remaining points are in the set U_0 and in positions in $[0,1]$.

All the constant approximations with 2 centers and 4 outliers have one center in $[0,1]$, one center between v_5, v_6, v_7 or v_8 and the outliers are v_1, v_2, v_3 and v_4. Instead all the constant approximations with 3 centers and 4 outliers have one center in $[0,1]$, one center between v_1 or v_2, one center between v_3 or v_4 and the outliers are v_5, v_6, v_7 and v_8. So if we do not use universal outliers a point between v_5, v_6, v_7 or v_8 is a center for the solution with 2 centers and an outlier for the solution for 3 centers, but this is impossible in hierarchical solution(refer to figure 2 for a visual representation). In fact, in hierarchical solution if a point is a center for

some k it has to remain center for all the larger values of k. Thus it is impossible to have a robust hierarchical solution without a set of universal outliers. □

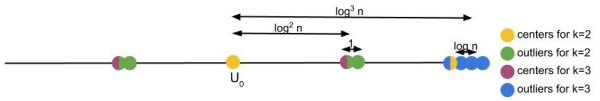

Figure 2: Visual representation for $k = 2,3$ and $z = 4$. The different set of outliers and centers are represented using different colors. Nodes with multiple colors are outliers or centers for multiple values of k.

4.2 Upper bound

In this section, we prove that the lower bound of Lemma 1 is asymptotically tight. In particular, we present a constructive proof for the existence of a set of universal outliers of size $O(z^2)$ that once removed allow the existence of algorithms that achieves an $O(1)$ approximation for hierarchical k-center. In particular, thanks to the result of Dasgupta and Long [8] on hierarchical k-center clustering it suffices to prove that there is a set S of size $O(z^2)$ such that for every k $\mathrm{OPT}_k(X \setminus S) \leq O(1) \cdot \mathrm{OPT}_k^z$.

Note that in this section our construction will not be algorithmic, in particular we assume to know the optimal solution and the optimal set of outliers, O_k, for each k. In the next section we will show how to turn this existential result into a polynomial time algorithm that identifies the set of universal outliers and produces the hierarchical clustering on the remaining points.

The main idea behind the construction is to include a node in the set of universal outliers only if by including it the optimal solution improves significantly and only if it does not have too many "close" points. More formally, the algorithm to find the universal set of outliers is shown in Algorithm 1.

Algorithm 1 Finding universal outliers

1: Let $0 < \alpha \leq 1/12$ be a parameter that we will choose later.
2: $t \leftarrow 0$
3: $k_0 \leftarrow 1$
4: $S_0 \leftarrow O_1$
5: **for** $2 \leq k \leq n$ **do**
6: **if** $\mathrm{OPT}_k^z < \alpha \cdot \mathrm{OPT}_k(X \setminus S_t)$ **then**
7: $t \leftarrow t + 1$
8: $k_t \leftarrow k$
9: $U_t \leftarrow S_{t-1} \cup O_k$
10: $A_t \leftarrow \left\{u \in U_t \mid B\left(u, 2 \cdot \mathrm{OPT}_k^z\right) \cap X \subseteq U_t\right\}$
11: $B_t \leftarrow \left\{u \in U_t \mid \left|B\left(u, 2 \cdot \mathrm{OPT}_k^z\right) \cap X\right| \leq z\right\}$
12: $S_t \leftarrow A_t \cap B_t$
13: **end if**
14: **end for**
15: $S \leftarrow S_t$

In the remaining of this section we prove that Algorithm 1 constructs a good universal set of outliers. More formally we show the following theorem:

Theorem 3. *For every metric space $\mathcal{M} = (X,d)$ with $|X| = n$ and $z > 0$ there exists a subset $S \subseteq X$ such that*

- $|S| \leq z^2$,
- *for every* $1 \leq k \leq n$, $\mathrm{OPT}_k(X \setminus S) \leq 28 \cdot \mathrm{OPT}_k^z$.

In order to prove the main theorem of the section we first show a series of technical lemma that we use to get the approximation guarantee. We start by showing that the cardinality of S is small:

Lemma 4. *For the resulting set S one has $|S| \leq z^2$.*

Proof. We prove that if for every t the size of S_t is at most z^2.

Consider the optimal k-center clustering of $X \setminus O_{k_t}$ (with all the clusters having radii at most $\mathrm{OPT}_{k_t}^z$). Consider all the clusters that consist of points from S_{t-1} exclusively and have cardinality at most z. Note that there are less than z such clusters otherwise, we could remove them and add all the points from O_{k_t} as singletons and this would contradict the invariant $\mathrm{OPT}_{k_t}^z < \alpha \cdot \mathrm{OPT}_{k_t}(X \setminus S_{t-1})$. So, in total there are at most $(z-1)z$ points in all these clusters. Moreover, all the points from S_{t-1} outside these clusters are filtered from S_t.

Thus, overall the total size of S_t is at most z^2. □

Now we focus on proving the approximation factor, the core idea of the proof is to show that $\mathrm{OPT}_{k_t}(X \setminus S_t)$ and $\mathrm{OPT}_{k_t}^z$ are a constant factor away and also $\mathrm{OPT}_{k_t}^z$ and $\mathrm{OPT}_{k_{t-1}}^z$ are a constant factor away and combine this two facts to prove the approximation. We start by introducing a lemma bounding the difference between the optimal k-center solutions on two set $P, Q \subseteq X$.

Lemma 5. *For every $P, Q \subseteq X$*

$$\mathrm{OPT}_k(P) \leq 2 \cdot \left(\mathrm{OPT}_k(Q) + \max_{p \in P} d(p,Q)\right).$$

Proof. First, note that

$$\mathrm{OPT}_k(P) \leq 2 \cdot \mathrm{OPT}_k(P \cup Q). \tag{1}$$

Indeed, consider the optimal k-center clustering for $P \cup Q$. First, we remove all the clusters that do not contain points from P. Second, in each of the remaining clusters we move the center to any point in P. Thus, the cost of the clustering can at most double.

Second, note that

$$\mathrm{OPT}_k(P \cup Q) \leq \mathrm{OPT}_k(Q) + \max_{p \in P} d(p,Q). \tag{2}$$

Indeed, consider the optimal k-center clustering for Q, attach all the points from $P \setminus Q$ to the closest clusters and apply the triangle inequality.

Finally, combining (1) and (2) we get the desired inequality. □

Using the previous Lemma we are now able to prove a relationship between $\mathrm{OPT}_{k_t}(X \setminus S_t)$ and $\mathrm{OPT}_{k_t}^z$.

Lemma 6. *For every t, $\mathrm{OPT}_{k_t}(X \setminus S_t) \leq 6 \cdot \mathrm{OPT}_{k_t}^z$.*

Proof. By Lemma 5 and the definition of O_{k_t}

$$\mathrm{OPT}_{k_t}(X \setminus S_t) \leq 2 \cdot \left(\mathrm{OPT}_{k_t}^z + \max_{u \in X \setminus S_t} d(u, X \setminus O_{k_t})\right).$$

Thus, it is sufficient to prove that for every $u \in X \setminus S_t$ one has $d(u, X \setminus O_{k_t}) \leq 2 \cdot \mathrm{OPT}_{k_t}^z$. If $u \notin O_{k_t}$, then the required distance is zero. So, we can assume wlog that $u \in O_{k_t} \setminus S_t$. Since $O_{k_t} \subseteq U_t$, u was filtered out in the line 9 of the code. This means that one of the two possibilities holds. The first is that $B\left(u, 2 \cdot \mathrm{OPT}_{k_t}^z\right) \cap X \not\subseteq U_t$, but since $O_{k_t} \subseteq U_t$, we have $d(u, X \setminus O_{k_t}) \leq 2 \cdot \mathrm{OPT}_{k_t}^z$. The second possibility is that $\left|B\left(u, 2 \cdot \mathrm{OPT}_{k_t}^z\right) \cap U_t\right| > z$. But since $|O_{k_t}| \leq z$ it again means that $d(u, X \setminus O_{k_t}) \leq 2 \cdot \mathrm{OPT}_{k_t}^z$. □

Now we focus on the relationship between $\mathrm{OPT}_{k_t}^z$ and $\mathrm{OPT}_{k_{t-1}}^z$.

LEMMA 7. *For every* $t \geq 1$, $\text{OPT}^z_{k_t} < 6\alpha \cdot \text{OPT}^z_{k_{t-1}}$.

PROOF.

$$\text{OPT}^z_{k_t} < \alpha \cdot \text{OPT}_{k_t}(X \setminus S_{t-1}) \leq \alpha \cdot \text{OPT}_{k_{t-1}}(X \setminus S_{t-1}) \leq 6\alpha \cdot \text{OPT}^z_{k_{t-1}}$$

The first step is due to the condition in the "for" loop. The last inequality is due to Lemma 6. □

After studying the relationship between $\text{OPT}_{k_t}(X \setminus S_t)$ and $\text{OPT}^z_{k_t}$ and between $\text{OPT}^z_{k_t}$ and $\text{OPT}^z_{k_{t-1}}$, we are now ready to prove that $\text{OPT}_{k^*}(X \setminus S)$ and $\text{OPT}^z_{k^*}$ are close.

LEMMA 8. *For every* $1 \leq k^* \leq n$

$$\text{OPT}_{k^*}(X \setminus S) \leq \left(\frac{2}{\alpha} + \frac{24\alpha}{1-6\alpha}\right) \cdot \text{OPT}^z_{k^*}.$$

Let λ be the largest integer such that $k_\lambda \leq k^*$. Let $\tilde{S} = S_\lambda$. Before showing the previous lemma, we prove two lemmas about $\tilde{S}$.

LEMMA 9.

$$\text{OPT}_{k^*}(X \setminus \tilde{S}) \leq \frac{1}{\alpha} \cdot \text{OPT}^z_{k^*}.$$

PROOF. If $k_\lambda < k^*$, then we are done due to the invariant we test in the "for" loop.

If $k_\lambda = k^*$, then by Lemma 6

$$\text{OPT}_{k^*}(X \setminus \tilde{S}) \leq 6 \cdot \text{OPT}^z_{k^*} \leq \frac{1}{\alpha} \cdot \text{OPT}^z_{k^*}.$$

The last inequality is true, because $\alpha \leq 1/12$. □

LEMMA 10. *For every* $u \in X \setminus S$

$$d(u, X \setminus \tilde{S}) \leq \frac{12\alpha}{1-6\alpha} \cdot \text{OPT}^z_{k^*}.$$

PROOF. If $u \in X \setminus \tilde{S}$, then the inequality is true. So we can assume that $u \in \tilde{S} \setminus S$. There exists $t > \lambda$ such that $u \in S_{t-1} \setminus S_t$. Let us prove that

$$d(u, X \setminus \tilde{S}) \leq 2 \cdot \sum_{i=\lambda+1}^{t} \text{OPT}^z_{k_i}. \tag{3}$$

Let us prove this inequality via induction on t. Since $u \in S_{t-1} \setminus S_t$, there are two possible explanations of this fact. First, it could be that $B\left(u, 2 \cdot \text{OPT}^z_{k_t}\right) \cap X \not\subseteq U_t$. If $B\left(u, 2 \cdot \text{OPT}^z_{k_t}\right) \cap X \not\subseteq \tilde{S}$, then (3) is true. So, we can focus on the case where $B\left(u, 2 \cdot \text{OPT}^z_{k_t}\right) \cap X \subseteq \tilde{S}$. Thus, there exists $u' \in \tilde{S} \setminus S_{t-1}$ such that $d(u,u') \leq 2 \cdot \text{OPT}^z_{k_t}$ (otherwise the point would be in S_t). But if is the case we can invoke the induction hypothesis and the triangle inequality to get equation (3). The second possibility is that $\left|B\left(u, 2 \cdot \text{OPT}^z_{k_t}\right) \cap U_t\right| > z$. But since $u \in \tilde{S}$, we have $\left|B\left(u, 2 \cdot \text{OPT}^z_{k_t}\right) \cap \tilde{S}\right| \leq z$ and so there is a point in $X \setminus \tilde{S}$ with at most distance $2 \cdot \text{OPT}^z_{k_t}$ from u. (Note that here we use that $t > \lambda$ and thus by Lemma 18 one has $2 \cdot \text{OPT}^z_{k_t} \leq 12\alpha \cdot \text{OPT}^z_{k_\lambda} \leq \text{OPT}_{k_\lambda}$ since $\alpha \leq 1/12$).

Finally, from (3), Lemma 7 and the fact that $k_{\lambda+1} > k^*$ we get

$$\begin{aligned} d(u, X \setminus \tilde{S}) &\leq 2 \cdot \text{OPT}^z_{k_{\lambda+1}} \cdot \sum_{i=0}^{\infty} (6\alpha)^i \leq \text{OPT}^z_{k_{\lambda+1}} \cdot \frac{2}{1-6\alpha} \\ &\leq \frac{12\alpha}{1-6\alpha} \cdot \text{OPT}^z_{k^*}. \end{aligned}$$

□

Using the previous two lemma we can prove Lemma 8.

PROOF. (of Lemma 8)

By combining Lemma 5, Lemma 9 and Lemma 10 we get

$$\begin{aligned} \text{OPT}_{k^*}(X \setminus S) &\leq 2 \cdot (\text{OPT}_{k^*}(X \setminus \tilde{S}) + \max_{u \in X \setminus S} d(u, X \setminus \tilde{S})) \\ &\leq \left(\frac{2}{\alpha} + \frac{24\alpha}{1-6\alpha}\right) \cdot OPT^z_{k^*}. \end{aligned}$$

□

We are now ready to show our main theorem in fact by using the result of Lemma 4 and Lemma 8 and by choosing choose $\alpha = 1/12$ we get Theorem 3.

5. HIERARCHICAL CLUSTERING WITH OUTLIERS

In this section we give a polynomial time algorithm that constructs a hierarchical clustering that uses a set S of universal outliers of size $O(z^2)$ and for every k gives a constant approximation to OPT^z_k.

The core idea of the algorithm is to use the same schema of our constructive proof for the universal set of outliers to find the outliers and then to apply a hierarchical k-center clustering on the remaining points. For $1 \leq k \leq n$ let O_k denote the set of z outliers for k-center clusterings of X computed by a 3-approximation algorithm. Note that O_k can be computed in polynomial time using the algorithm from [4]. Let $\text{cost}_k(U)$ be the cost of a k-center clustering of $U \subseteq X$, found by the farthest-point traversal. To find S we now use Algorithm 2.

Algorithm 2 Finding a hierarchical universal k-center

```
1: t ← 0
2: k_0 ← 1
3: S_0 ← O_1
4: for 2 ≤ k ≤ n do
5:     if cost_k(X \ O_k) < α · cost_k(X \ S_t) then
6:         t ← t + 1
7:         k_t ← k
8:         U_t ← S_{t-1} ∪ O_k
9:         A_t ← {u ∈ U_t | B(u, 2 · cost_k(X \ O_k)) ∩ X ⊆ U_t}
10:        B_t ← {u ∈ U_t | |B(u, 2 · cost_k(X \ O_k)) ∩ X| ≤ z}
11:        S_t ← A_t ∩ B_t
12:    end if
13: end for
14: S ← S_t
```

Using the previous algorithm we can show that:

THEOREM 11. *For every metric space* $\mathcal{M} = (X,d)$ *with* $|X| = n$ *and* $z > 0$ *there exists a subset* $S \subseteq X$

- $|S| \leq z^2$,
- *for every* $1 \leq k \leq n$, $\text{OPT}_k(X \setminus S) \leq O(1) \cdot \text{OPT}^z_k$.
- *S can be found in polynomial time*

The full proof of Theorem 11 is given in Appendix A. The proof follows the lines of Theorem 3. The blow up in the approximation factor is given by the use of the 3-approximation polynomial time algorithm from [4] to compute a set O_k of outliers and by the 2-approximation farthest-first traversal polynomial time algorithm for computing $\text{cost}_k(X \setminus O_k)$.

Note that once we have found the universal set of outliers we can use the hierarchical k-center algorithm of Dasgupta and Long [8] on the remaining points to get a hierarchical k-center clustering with z outliers that uses a set of universal outliers of size $O(z^2)$. More formally:

COROLLARY 12. *There is a polynomial algorithm for the hierarchical k-center clustering with z outliers that uses a set S of universal outliers of size $O\left(z^2\right)$ and that compute a set of centers $c_1, c_2, \ldots, c_n$ and a function π such that, for every $1 \leq k \leq n$, $C(c_1, c_2, \ldots, c_k, \pi, X \setminus S) \leq O(1) \cdot \mathrm{OPT}_k^z$.*

6. STOCHASTIC HIERARCHICAL CLUSTERING

We define the *stochastic* variant of hierarchical k-center with outliers: find a hierarchical k-center solution on the metric space $\mathcal{M} = (X, d)$ which is good on average for any subset of h randomly selected points. Unfortunately this goal cannot always be achieved, for instance if the metric space contains $k+1$ points that are far away from each other.

For this reason, it is important to allow the presence of outliers also in the stochastic setting. [1] More precisely, we answer the following question: does it exist a small set of outliers that allows to compute a good *stochastic* and *hierarchical* k-center solution for a range of values $k = 1, \ldots, K$? We answer this question by finding a set S of outliers such that the expectation, over a random subset of h points from X, of the cost of an a-priori hierarchical clustering of X is for any fixed $k \in \{1, \ldots, K\}$ close to the expected cost of the optimal k-clustering. We show the following theorem:

THEOREM 13. *There exists a 48-approximation algorithm for* stochastic hierarchical *k-center on inputs of size h and maximum number of centers equal to K that uses at most $z = \frac{2K(K-1)\ln n}{h} n$ outliers.*

We remark that the solution computed on $X \setminus S$ is an a-priori solution which is good on average for any subset of h points sampled from X.

To show our Theorem we first prove some technical lemmas. We start by proving that there exists a set of outliers S of small size such that there exists a stochastic k-center solution for $X \setminus S$ that is on average also a good solution for $Y \setminus S$, where Y is randomly selected set of h points from X (possibly with repetitions). Initially we focus on the setting where k is fixed then we generalize the result for k between $1, 2, \ldots, K$.

Let us first fix a value of k. Let us also denote by $OPT_k^{X \setminus S}(Y \setminus S)$ the cost of the optimal k-center solution computed for $X \setminus S$ when applied to set $Y \setminus S$. We also denote by $OPT_k(Y)$ the cost of the optimal k-center solution computed for Y.

The goal is to determine a set of outliers S of minimum cardinality such that:

$$E_{Y \in X^h}[\mathrm{OPT}_k^{X \setminus S}(Y \setminus S)] / E_{Y \in X^h}[\mathrm{OPT}_k(Y)] = O(1).$$

Let $r = E_{Y \in X^h} OPT_k(Y)$ the expected cost of the k-center solution on a random set Y of size h. In the following we prove a claim on the existence of a set of k balls of radius $2r$ that cover most of the points of the metric space. This lemma will provide us with an upper bound on the size of the set of outliers that is needed in order to obtain a good *stochastic* solution. Note in fact that the existence of a good solution for $X \setminus S$ implies the existence of a good solution for $Y \setminus S$ since $Y \subseteq X$. A similar lemma is also proved in [12] for the universal stochastic set cover problem.

LEMMA 14. *There exists a set of k balls of radius $2r$ that cover all but a fraction $\delta = \frac{2k \log n}{h}$ of X.*

PROOF. Since the expected optimal cost is r, we know that at least $\frac{1}{2} n^h$ input instances, i.e., half of the input instances, will lead to a solution of radius at most $2r$. Moreover, there exists at most $p = n^k = e^{klnn}$ solutions of k balls of radius $2r$. Denote by X_i the set of points covered from solution i within radius $2r$. We prove that there must exist at least one solution X_i such that $|X_i| \geq (1-\delta)n$ for $\delta = \frac{2k \log n}{h}$. In order to prove this claim, assume by contradiction that all the solutions X_i have $|X_i| < (1-\delta)n \leq n e^{-\delta}$.

All the subsets of h points with a solution of radius at most $2r$ must have been selected among all subsets of h points covered from one of the solutions X_i. We therefore have the following inequality:

$$\sum_{i=1}^{p} |X_i|^h \geq \frac{1}{2} n^h.$$

Since there are at most e^{klnn} solutions, and by contradiction $|X_i| < n e^{-\delta}$ for all solutions, we obtain $e^{klnn}(n^{-\delta})^h \geq \frac{1}{2} n^h$ and therefore $e^{k \ln n - h\delta} \geq \frac{1}{2}((1-\delta)n)^h$. This is leads to a contradiction if $\delta \geq 2\frac{k \ln n}{h}$ since

$$e^{-k \ln n} < \frac{1}{2} n^h,$$

for $k, n, h > 1$. □

The previous lemma allows to claim the existence of set S of $z = \frac{2k \ln n}{h} n$ outliers such that a solution for k-center on $X \setminus S$ of value $2r$ is also a solution a good *stochastic solution* for $Y \setminus S$ since $Y \subseteq X$.

Now for a fixed k, we compute a 3-approximation solution [4] with set S of

$$z = \frac{2k \ln n}{h} n$$

outliers on metric space X. The returned solution provides a k-center solution to be used for any randomly chosen subset Y of h points with set of outliers S. Given Lemma 14, this algorithm is a 6 approximation of the expected cost of a k-center for a random subset Y of h randomly chosen points of the metric space. This solution is meaningful as long as as

$$h = \Omega(k \ln n).$$

We therefore conclude with the following:

LEMMA 15. *There exists a 6-approximation algorithm for k-center for fixed k and sample size h that uses at most $z = \frac{2k \ln n}{h} n$ outliers.*

We now extend this solution to a *stochastic hierarchical* k-center problem with outliers for a number of centers in a range $k = 1, \ldots, K$. We take the union of the set of outliers for each value of $k = 1, \ldots, K$. We obtain a total number of outliers equal to

$$\frac{2K(K-1)\ln n}{h} n$$

[1] Note that in order to have a hierarchical clustering we need to have an assignment of all points to their centers and this assignment cannot be based only on d [8] so we need to assume to know X in advance.

that is meaningful as long as

$$h = \Omega(K(K-1)\ln n).$$

The algorithm for the *stochastic* variant of hierarchical k-center is as follows:

We run an 8-approximation hierarchical algorithm [8] for the whole metric space after we remove a set of universal outliers of cardinality $\frac{2K(K-1)\ln n}{h} n$(we remove $z = \frac{2k\ln n}{h} n$ outliers for each k as in the previous algorithm), note that by removing the universal set of outliers we loose at most a factor of 3 in the optimal solution. This algorithm provides a 24-approximation of the optimal solution $\mathrm{OPT}_k^S(X)$ that we know it is at most equal to $2r = 2E_Y[OPT_k(Y)]$ for input set Y of cardinality h. Thus we get the main theorem of this section Theorem 13.

7. REFERENCES

[1] Anagnostopoulos, A., Grandoni, F., Leonardi, S., and Sankowski, P. Online network design with outliers. In *Proceedings of the 37th International Colloquium Conference on Automata, Languages and Programming* (Berlin, Heidelberg, 2010), ICALP'10, Springer-Verlag, pp. 114–126.

[2] Balcan, N., and Gupta, P. Robust hierarchical clustering. In *COLT* (2010).

[3] Bryant, D., and Berry, V. A structured family of clustering and tree construction methods. *Advances in Applied Mathematics 27*, 4 (2001), 705 – 732.

[4] Charikar, M., Khuller, S., Mount, D. M., and Narasimhan, G. Algorithms for facility location problems with outliers. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms* (Philadelphia, PA, USA, 2001), SODA '01, Society for Industrial and Applied Mathematics, pp. 642–651.

[5] Charikar, M., O'Callaghan, L., and Panigrahy, R. Better streaming algorithms for clustering problems. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, STOC03* (2003), pp. 30–39.

[6] Cheng, D., Kannan, R., Vempala, S., and Wang, G. A divide-and-merge methodology for clustering. *ACM Trans. Database Syst. 31*, 4 (Dec. 2006), 1499–1525.

[7] Das, A., and Kenyon, C. On hierarchical diameter-clustering and the supplier problems. In *In Proc. WAOAÕ06, 4th Workshop on Approximation and Online Algorithms, September 2006* (2006), Springer.

[8] Dasgupta, S., and Long, P. M. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences 70*, 4 (2005), 555–569.

[9] Garg, N., Gupta, A., Leonardi, S., and Sankowski, P. Stochastic analyses for online combinatorial optimization problems. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Philadelphia, PA, USA, 2008), SODA '08, Society for Industrial and Applied Mathematics, pp. 942–951.

[10] Gollapudi, S., Kumar, R., and Sivakumar, D. Programmable clustering. In *Proceedings of the Twenty-fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (New York, NY, USA, 2006), PODS '06, ACM, pp. 348–354.

[11] Gonzalez, T. F. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science 38*, 0 (1985), 293 – 306.

[12] Grandoni, F., Gupta, A., Leonardi, S., Miettinen, P., Sankowski, P., and Singh, M. Set covering with our eyes closed. *SIAM J. Comput. 42*, 3 (2013), 808–830.

[13] Guha, S., Rastogi, R., and Shim, K. Cure: an efficient clustering algorithm for large databases. In *Proceedings of the 24th Annual ACM SIGMOD* (1998).

[14] Hochbaum, D. S., and Shmoys, D. B. A best possible heuristic for the k-center problem. *Mathematics of operations research 10*, 2 (1985), 180–184.

[15] Jain, A., Murty, M., and Flynn, P. Data clustering: a review. *ACM computing surveys* (1999).

[16] Jain, A. K., and Dubes, R. C. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[17] Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: A review. *ACM Comput. Surv. 31*, 3 (Sept. 1999), 264–323.

[18] Johnson., S. C. Hierarchical clustering schemes. *Psychometrika* (1967).

[19] McCutchen, R. M., and Khuller, S. Streaming algorithms for k-center clustering with outliers and with anonymity. In *Proceedings of the 11th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX '2008)* (2008), pp. 165–178.

[20] Narasimhan, M., Jojic, N., and Bilmes, J. Q-clustering. In *Advances in Neural Information Processing Systems (NIPS)* (2006), pp. 348–354.

[21] Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* (1963).

[22] Wishart., D. Mode analysis: a generalization of nearest neighbour which reduces chaining effects. *Numerical Taxonomy* (1969).

APPENDIX

A. PROOFS OF SECTION 5

Here we give the proof of Theorem 11, before proving it we restate it:

Theorem 16. *For every metric space $\mathcal{M} = (X,d)$ with $|X| = n$ and $z > 0$ there exists a subset $S \subseteq X$*

- $|S| \le z^2$,
- *for every* $1 \le k \le n$, $\mathrm{OPT}_k(X \setminus S) \le O(1) \cdot \mathrm{OPT}_k^z$.
- *S can be found in polynomial time*

The proof follows from the same scheme of the techniques of the constructive proof, the main difference is that here we will handle suboptimal solutions and outliers.

Proof. Let $0 < \alpha \le 1/360$ be a parameter that we will choose later.

Lemma 17. *For every t,* $\mathrm{OPT}_{k_t}(X \setminus S_t) \le 30 \cdot \mathrm{OPT}_{k_t}^z$.

Proof. By Lemma 5 and the definition of O_{k_t}

$$\mathrm{OPT}_{k_t}(X \setminus S_t) \le 2(3 \cdot \mathrm{OPT}_{k_t}^z + \max_{u \in X \setminus S_t} d(u, X \setminus O_{k_t})).$$

Thus, it is sufficient to prove that for every $u \in X \setminus S_t$ one has $d(u, X \setminus O_{k_t}) \le 12 \cdot \mathrm{OPT}_{k_t}^z$. If $u \notin O_{k_t}$, then the required distance is zero. So, we can assume wlog that $u \in O_{k_t} \setminus S_t$. Since $O_{k_t} \subseteq U_t$, u was filtered out in the line 9 of the code. This means that one of the two possibilities holds. The first is that $B(u, 2 \cdot \mathrm{cost}_{k_t}(X \setminus O_{k_t})) \cap X \not\subseteq U_t$, but since $O_{k_t} \subseteq U_t$, we have

$$d(u, X \setminus O_{k_t}) \le 2 \cdot \mathrm{cost}_{k_t}(X \setminus O_{k_t}) \le 4 \cdot \mathrm{OPT}_{k_t}(X \setminus O_{k_t}) \le 12 \cdot \mathrm{OPT}_{k_t}^z.$$

The inequality above follows since $\text{cost}_{k_t}(X \setminus O_{k_t})$ is computed by a 2-approximation k-center algorithm whereas O_{k_t} is computed with a 3-approximation algorithm for k-center with outliers.

The second possibility is that $|B(u, 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})) \cap U_t| > z$. But since $|O_{k_t}| \le z$ it again means that $d(u, X \setminus O_{k_t}) \le 12 \cdot \text{OPT}^z_{k_t}$. □

Lemma 18. *For every $t \ge 1$, $\text{OPT}^z_{k_t} < 60\alpha \cdot \text{OPT}^z_{k_{t-1}}$.*

Proof.

$$\begin{aligned} \text{OPT}^z_{k_t} \le \text{cost}_{k_t}(X \setminus O_{k_t}) &< \alpha \cdot \text{cost}_{k_t}(X \setminus S_{t-1}) \\ &\le 2\alpha \cdot \text{OPT}_{k_t}(X \setminus S_{t-1}) \\ &\le 2\alpha \cdot \text{OPT}_{k_{t-1}}(X \setminus S_{t-1}) \\ &\le 60\alpha \cdot \text{OPT}^z_{k_{t-1}}. \end{aligned}$$

The second step is due to the condition in the "for" loop. The last inequality is due to Lemma 17. □

Now we prove that the resulting set S is relatively good for every k.

Lemma 19. *For every $1 \le k^* \le n$*

$$\text{OPT}_{k^*}(X \setminus S) \le 2163 \cdot \text{OPT}^z_{k^*}.$$

Proof. Let λ be the largest integer such that $k_\lambda \le k^*$. Let $\tilde{S} = S_\lambda$.

We prove two lemmas about $\tilde{S}$.

Lemma 20. $\text{OPT}_{k^*}(X \setminus \tilde{S}) \le \frac{6}{\alpha} \cdot \text{OPT}^z_{k^*}$.

Proof. If $k_\lambda < k^*$, then we are done due to the invariant we test in the "for" loop. Indeed,

$$\begin{aligned} \text{OPT}_{k^*}(X \setminus \tilde{S}) &\le \text{cost}_{k^*}(X \setminus \tilde{S}) \le \frac{1}{\alpha} \cdot \text{cost}_{k^*}(X \setminus O_{k^*}) \\ &\le \frac{2}{\alpha} \cdot \text{OPT}_{k^*}(X \setminus O_{k^*}) \le \frac{6}{\alpha} \cdot \text{OPT}^z_{k^*}. \end{aligned}$$

If $k_\lambda = k^*$, then by Lemma 17

$$\text{OPT}_{k^*}(X \setminus \tilde{S}) \le 30 \cdot \text{OPT}^z_{k^*} \le \frac{6}{\alpha} \cdot \text{OPT}^z_{k^*}.$$

The last inequality is true, because $\alpha < 1/5$. □

Lemma 21. *For every $u \in X \setminus S$*

$$d(u, X \setminus \tilde{S}) \le \frac{720}{1 - 60\alpha} \cdot \text{OPT}^z_{k^*}.$$

Proof. If $u \in X \setminus \tilde{S}$, then the inequality is true. So we can assume that $u \in \tilde{S} \setminus S$. There exists $t > \lambda$ such that $u \in S_{t-1} \setminus S_t$. Let us prove that

$$d(u, X \setminus \tilde{S}) \le 2 \cdot \sum_{i=\lambda+1}^{t} \text{cost}_{k_i}(X \setminus O_{k_i}). \tag{4}$$

Let us prove this inequality via induction on t. Since $u \in S_{t-1} \setminus S_t$, there are two possible explanations of this fact. First, it could be that $B(u, 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})) \cap X \not\subseteq U_t$. If $B(u, 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})) \cap X \not\subseteq \tilde{S}$, then (4) is true. So, we can assume that $B(u, 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})) \cap X \subseteq \tilde{S}$. Thus, there exists $u' \in \tilde{S} \setminus S_{t-1}$ such that $d(u, u') \le 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})$. In this case we can invoke the induction hypothesis and the triangle inequality. The second possibility is that $|B(u, 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})) \cap U_t| > z$. Note that since $u \in \tilde{S}$ and thus $|B(u, 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})) \cap \tilde{S}| \le z$. (Note that here we use that $t > \lambda$ and thus by Lemma 18 one has

$$2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t}) < 12 \cdot \text{OPT}^z_{k_t} \le 720\alpha \text{OPT}^z_{k_\lambda}$$

$$\le 720\alpha \cdot \text{cost}_{k_\lambda}(X \setminus O_{k_\lambda}) \le 2 \cdot \text{cost}_{k_\lambda}(X \setminus O_{k_\lambda}),$$

since $\alpha \le 1/360$. So in this case $d(u, X \setminus \tilde{S}) \le 2 \cdot \text{cost}_{k_t}(X \setminus O_{k_t})$.

Finally, from (4), Lemma 18 and the fact that $k_{\lambda+1} > k^*$ we get

$$d(u, X \setminus \tilde{S}) \le 12 \cdot \text{OPT}^z_{k_{\lambda+1}} \cdot \sum_{i=0}^{\infty} (60\alpha)^i \le \text{OPT}^z_{k_{\lambda+1}} \cdot \frac{12}{1 - 60\alpha}$$

$$\le \frac{720\alpha}{1 - 60\alpha} \cdot \text{OPT}^z_{k^*}.$$

□

Finally, combining Lemma 5, Lemma 20 and Lemma 21 we get

$$\text{OPT}_{k^*}(X \setminus S) \le \text{OPT}_{k^*}(X \setminus \tilde{S}) + \max_{u \in X \setminus S} d(u, X \setminus \tilde{S})$$

$$\le \left(\frac{6}{\alpha} + \frac{720\alpha}{1 - 60\alpha} \right) \cdot OPT^z_{k^*}.$$

We can choose $\alpha = 1/360$ to get a 2163-approximation.

Finally we prove that S is relatively small.

Lemma 22. *For the resulting set S one has $|S| \le z^2$.*

Proof. We prove that if for every t the size of S_t is at most z^2.

Consider a k-center clustering of $X \setminus O_{k_t}$ with all the clusters having radii at most $\text{cost}_{k_t}(X \setminus O_{k_t})$. Consider all the clusters that consist of points from S_{t-1} exclusively and have cardinality at most z. Clearly, there are less than z such clusters. Indeed, otherwise we could upper bound

$$\text{OPT}_{k_t}(X \setminus S_{t-1}) \le \text{cost}_{k_t}(X \setminus O_{k_t})$$

by removing these clusters and attaching points from $O_{k_t} \setminus S_{t-1}$ as singletons. But since

$$\text{OPT}_{k_t}(X \setminus S_{t-1}) \ge \frac{1}{2} \cdot \text{cost}_{k_t}(X \setminus S_{t-1}),$$

we would get

$$\text{cost}_{k_t}(X \setminus O_{k_t}) \ge \frac{1}{2} \cdot \text{cost}_{k_t}(X \setminus S_{t-1}),$$

which contradicts the invariant

$$\text{cost}_{k_t}(X \setminus O_{k_t}) < \alpha \cdot \text{cost}_{k_t}(X \setminus S_{t-1})$$

(here we use that $\alpha \le 1/2$). So, in total there are at most $(z-1)z$ points in all these clusters. Moreover, all the points from S_{t-1} outside these clusters are filtered from S_t.

Thus, overall the total size of S_t is at most z^2. □